Santa Fe Institute Working Paper 10-11-029
arxiv.org:arXiv:1011.1581 [nlin.CD]

Asymptotic Synchronization for Finite-State Sources

Nicholas F. Travers[1, 2, *] and James P. Crutchfield[1, 2, 3, 4, †]

[1]*Complexity Sciences Center*
[2]*Mathematics Department*
[3]*Physics Department*
University of California at Davis,
One Shields Avenue, Davis, CA 95616
[4]*Santa Fe Institute*
1399 Hyde Park Road, Santa Fe, NM 87501

(Dated: January 5, 2011)

We extend a recent synchronization analysis of exact finite-state sources to nonexact sources for which synchronization occurs only asymptotically. Although the proof methods are quite different, the primary results remain the same. We find that an observer's average uncertainty in the source state vanishes exponentially fast and, as a consequence, an observer's average uncertainty in predicting future output converges exponentially fast to the source entropy rate.

PACS numbers: 02.50.-r 89.70.+c 05.45.Tp 02.50.Ey

I. INTRODUCTION

In Ref. [1] we analyzed the synchronization process for exact ϵ-machines, where the observer may come to know the internal state of the machine with certainty after only a finite number of measurements. Here, we examine the case of nonexact ϵ-machines, where the observer may only synchronize to the machine's state asymptotically. Although the analysis differs, the behavior is qualitatively similar to the exact case in the sense that an observer (on average) synchronizes to a nonexact machine exponentially fast. That is, there exist constants $K > 0$ and $0 < \alpha < 1$ such that the average state entropy $\mathcal{U}(L) \leq K\alpha^L$, for all $L \in \mathbb{N}$.

Our development is organized as follows. Section II briefly reviews the synchronization problem and provides the essential definitions for our results. Section III presents an intuitive picture of the synchronization process, using it to derive a formula for $\phi(w)$, the conditional state distribution induced by a word w. Section IV establishes a formula for the entropy rate of a finite-state ϵ-machine. Section V uses the entropy-rate formula to prove the existence of averaged asymptotic synchronization. Section VI builds on this result to prove our main theorem—the Nonexact Machine Synchronization Theorem. Section VII uses this theorem to show that, for any nonexact ϵ-machine, the state entropy $\mathcal{U}(L)$ vanishes exponentially fast and the length-L entropy-rate approximation $h_\mu(L)$ converges exponentially fast to the machine's entropy rate. Finally, Sec. VIII summarizes our results and examines possible extensions.

*Electronic address: ntravers@math.ucdavis.edu
†Electronic address: chaos@cse.ucdavis.edu

II. BACKGROUND

This section lays out the necessary definitions and background for our results. For a more thorough introduction the reader is referred to Ref. [1], where a similar but more detailed presentation is given.

A. Machines

Definition 1. Hidden Markov machine*: A* finite-state edge-label hidden Markov machine (HMM) *consists of*

1. *a finite set of states* $\boldsymbol{\mathcal{S}} = \{\sigma_1, ..., \sigma_N\}$,
2. *a finite alphabet of symbols* $\mathcal{A}$, *and*
3. *a set of* N *by* N *symbol-labeled transition matrices* $T^{(x)}$, $x \in \mathcal{A}$, *where* $T^{(x)}_{ij}$ *is the probability of transitioning from state* σ_i *to state* σ_j *on symbol* x. *The corresponding internal state-to-state transition matrix is denoted* $T = \sum_{x\in\mathcal{A}} T^{(x)}$.

A hidden Markov machine can be depicted as a directed graph with labeled edges. The nodes are the states $\{\sigma_1, ..., \sigma_N\}$ *and for all* x, i, j *with* $T^{(x)}_{ij} > 0$, *there is an edge from state* σ_i *to state* σ_j *labeled* $p|x$ *for the symbol* x *and transition probability* $p = T^{(x)}_{ij}$. *We require that the transition matrices* $T^{(x)}$ *be such that this graph is strongly connected.*

A hidden Markov machine M generates a stationary process $\mathcal{P} = (X_L)_{L\geq 0}$ as follows. Initially, M starts in some state σ_{i^*} chosen according to the stationary distribution π over machine states—the distribution satisfying $\pi T = \pi$. It then picks an outgoing edge according to their relative transition probabilities $T^{(x)}_{i^*j}$, emits the symbol x^* labeling this edge, and follows the edge to a new state σ_{j^*}. The next output symbol and state are consequently chosen in a similar fashion, and this procedure is repeated indefinitely.

We denote by $\mathcal{S}_0, \mathcal{S}_1, \mathcal{S}_2, \ldots$ the random variables (RVs) for the sequence of machine states visited and

by $X_0, X_1, X_2, \ldots$ the RVs for the associated sequence of output symbols generated. The sequence of states $(\mathcal{S}_L)_{L\geq 0}$ is a Markov chain with transition kernel T. However, the stochastic process we consider is not the sequence of states, but rather the associated sequence of outputs $(X_L)_{L\geq 0}$, which is not normally Markov. We assume that an observer of the process sees the sequence of outputs, but does not have direct access to the machine's "hidden" internal states.

Example: Even Process Machine Figure 1 gives an HMM for the Even Process, a machine that has been studied extensively [2]. Its name derives from the feature that in its output there are always an even number of 1s between consecutive 0s. The transition matrices are:

$$T^{(0)} = \begin{pmatrix} p & 0 \\ 0 & 0 \end{pmatrix} ,$$
$$T^{(1)} = \begin{pmatrix} 0 & 1-p \\ 1 & 0 \end{pmatrix} .$$

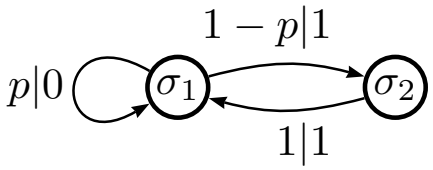

FIG. 1: A hidden Markov machine (the ϵ-machine) for the Even Process. The transitions denote the probability p of generating symbol x as $p|x$.

The following notation will be used for sequences of output RVs:

1. $\overrightarrow{X} = X_0 X_1 \ldots$,
2. $\overrightarrow{X}^L = X_0 X_1 \ldots X_{L-1}$, and
3. $\overrightarrow{X}_t^L = X_t X_{t+1} \ldots X_{t+L-1}$.

Definition 2. *A* finite-state ϵ-machine *is a finite-state edge-label hidden Markov machine with the following properties:*

1. Unifilarity*: For each state $\sigma_k \in \boldsymbol{\mathcal{S}}$ and each symbol $x \in \mathcal{A}$ there is at most one outgoing edge from state σ_k labeled with symbol x.*
2. Probabilistically distinct states*: For each pair of distinct states $\sigma_k, \sigma_j \in \boldsymbol{\mathcal{S}}$ there exists some finite word $w = x_0 x_1 \ldots x_{L-1}$ such that:*

$$\Pr(\overrightarrow{X}^L = w | \mathcal{S}_0 = \sigma_k) \neq \Pr(\overrightarrow{X}^L = w | \mathcal{S}_0 = \sigma_j) .$$

Example (continued) The Even Process machine given above is also an ϵ-machine. It is clearly unifilar, and σ_1 can generate the symbol 0 whereas σ_2 cannot, so the states are probabilistically distinct.

Remark. *ϵ-Machines were originally defined in Ref. [3] as hidden Markov machines whose states, known as* causal states*, were the equivalence classes of infinite pasts $\overleftarrow{x}$ with the same probability distribution over futures $\overrightarrow{x}$. This "history machine" definition is, in fact, equivalent to the "generating machine" definition presented above in the finite-state case. Although, this is not immediately apparent. Formally, it follows from the synchronization results established here and in Ref. [1].*

We now provide the definitions for two extensions of an ϵ-machine M that are necessary for our proofs later on: the edge machine M_{edge} and the power machine M^n. In what follows:

1. $\Pr(x|\sigma_k) \equiv \Pr(X_0 = x | \mathcal{S}_0 = \sigma_k)$,
2. $\Pr(w|\sigma_k) \equiv \Pr(\overrightarrow{X}^{|w|} = w | \mathcal{S}_0 = \sigma_k)$,
3. $I(x, k, j)$ denotes the indicator function of the transition from state σ_k to state σ_j on symbol x, and
4. $I(w, k, j)$ denotes the indicator function of the transition from state σ_k to state σ_j on the word w.

That is, $I(x, k, j) = 1$ if $\sigma_k \stackrel{x}{\to} \sigma_j$ and 0 otherwise; $I(w, k, j) = 1$ if $\sigma_k \stackrel{w}{\to} \sigma_j$ and 0 otherwise.

Definition 3. *For an ϵ-machine M, the corresponding* edge machine *M_{edge} is the Markov chain whose states are the outgoing edges of M. That is, the states are the pairs (x, σ_k) such that $\Pr(x|\sigma_k) > 0$, and the transition probabilities are defined as:*

$$\Pr((x, \sigma_k) \to (y, \sigma_j)) = \Pr(y|\sigma_j) I(x, k, j) .$$

A sequence of M_{edge} states visited by the Markov chain corresponds to a sequence of edges visited by the original machine M. The process $\mathcal{P}_{edge}$ generated by M_{edge} can be thought of as the *bi-process* $(X_L, \mathcal{S}_L)_{L\geq 0}$ generated by the original machine M as it moves from state to state generating symbols. Note that since M's graph is strongly connected, M_{edge}'s graph is as well. Hence, the edge-label Markov chain is irreducible and has a unique stationary distribution π_{edge}. See Fig. 2(top).

Definition 4. *Let M be an ϵ-machine, and let n be relatively prime to the period p of M's graph. The* power machine *M^n is defined to be the ϵ-machine with the states of M, output symbols which are length-n words generated by M, and transition probabilities given by:*

$$\Pr(\sigma_k \stackrel{w}{\to} \sigma_j) = \Pr(w|\sigma_k) I(w, k, j) .$$

The power machine M^n generates the same process as the original machine M, but over length-n blocks.

Note that since M is by definition unifilar with probabilistically distinct states, M^n is also necessarily unifilar with probabilistically distinct states. Furthermore, it can be shown that for n relatively prime to $p = \text{period}(M)$ the graph of M^n is strongly connected. Therefore, for n relatively prime to p, M^n is indeed an ϵ-machine for the process $\mathcal{P}^n$. See Fig. 2(bottom).

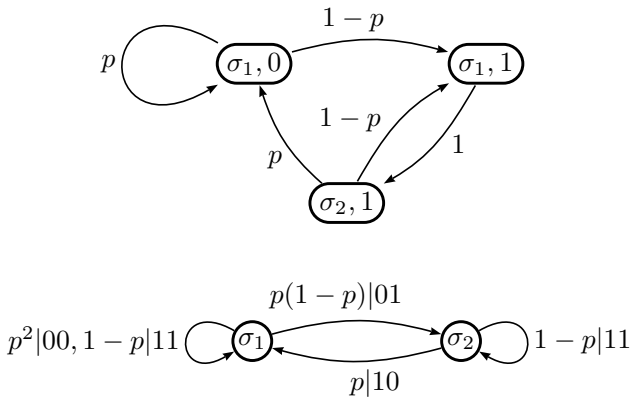

FIG. 2: Examples of M_{edge} (top) and M^2 (bottom) for the Even Process ϵ-machine M.

Definition 5. *For an ϵ-machine M the* minimum distinguishing length L^* *is the shortest length L such that the probability distributions over futures $\overrightarrow{X}^L$ of length L are distinct for each pair of distinct states σ_k and σ_j:*

$$L^* \equiv \min\{L : \Pr(\overrightarrow{X}^L|\mathcal{S}_0 = \sigma_k) \neq \Pr(\overrightarrow{X}^L|\mathcal{S}_0 = \sigma_j) \text{ , for all } k \neq j\} \ .$$

If a machine M has a minimum distinguishing length L^*, we also say that M has length-L^* future distinguishable states.

Note that L^* must be finite for any ϵ-machine, since ϵ-machines have probabilistically distinct states, and that, for $n \geq L^*$ and relatively prime to $p = \text{period}(M)$, M^n is an ϵ-machine with a minimum distinguishing length of 1.

B. Synchronization

Although we assume that our observer is not able to directly see the ϵ-machine's internal state ($\mathcal{S}_L$), it is able to see the output symbols generated by the machine (the X_L's). Thus, the observer may attempt to infer the internal machine state through observations of the output. We are interested in studying the procedure by which the observer synchronizes to the machine's state through these observations. Due to unifilarity, we know that if an observer is able to completely synchronize to the machine's internal state at some time $T > 0$, it remains synchronized for all future times $T' \geq T$. For simplicity, we assume that the initial state is chosen according to the stationary distribution π, so that the process generated by the machine is stationary, and also that the observer has knowledge of this fact.

For a word w of length L generated by the machine let $\phi(w) \equiv \Pr(\boldsymbol{\mathcal{S}}|w)$ be the observer's *belief distribution* as to the current state of the machine after observing w. That is,

$$\begin{aligned}\phi(w)_k &= \Pr(\mathcal{S}_L = \sigma_k|\overrightarrow{X}^L = w) \\ &\equiv \Pr(\mathcal{S}_L = \sigma_k|\overrightarrow{X}^L = w, \mathcal{S}_0 \sim \pi) \ .\end{aligned}$$

And, define the observer's uncertainty in the machine state after observing w as:

$$\begin{aligned}u(w) &= H[\phi(w)] \\ &= H[\mathcal{S}_L|\overrightarrow{X}^L = w] \ .\end{aligned}$$

Let $\mathcal{L}(M)$ denote the set of all finite words that M can generate, $\mathcal{L}_L(M)$ the set of all length-L words it can generate, and $\mathcal{L}_\infty(M)$ the set of all infinite sequences $\overrightarrow{x} = x_0x_1...$ that it can generate.

Definition 6. *A word $w \in \mathcal{L}(M)$ is a* synchronizing word *(or* sync word*) for M if $u(w) = 0$; that is, if the observer knows the current state of the machine with certainty after observing w.*

We denote the set of M's infinite synchronizing sequences as $\text{SYN}(M)$ and the set of M's infinite weakly synchronizing sequences as $\text{WSYN}(M)$:

$$\text{SYN}(M) = \{\overrightarrow{x} \in \mathcal{L}_\infty(M) : u(\overrightarrow{x}^L) = 0 \text{ for some } L\}, \text{ and}$$
$$\text{WSYN}(M) = \{\overrightarrow{x} \in \mathcal{L}_\infty(M) : u(\overrightarrow{x}^L) \to 0 \text{ as } L \to \infty\} \ .$$

Definition 7. *An ϵ-machine M is* exactly synchronizable *(or simply* exact*) if $\Pr(\text{SYN}(M)) = 1$; that is, if the observer synchronizes to almost every (a.e.) sequence the machine generates in finite time.*

Definition 8. *An ϵ-machine M is* asymptotically synchronizable *if $\Pr(\text{WSYN}(M)) = 1$; that is, if the observer's uncertainty in the machine state vanishes asymptotically for a.e. sequence the machine generates.*

Examples:

- The Even Process ϵ-machine is an exact machine. Any word containing a 0 is a sync word for this machine, and almost every $\overrightarrow{x}$ generated by this machine contains at least one 0.
- The ABC machine (Fig. 3) is not exactly synchronizable, but it is asymptotically synchronizable.

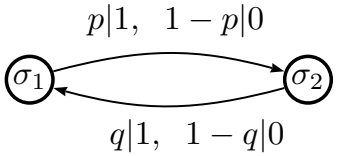

FIG. 3: The Alternating Biased Coin (ABC) machine: The process it generates can be thought of as alternately flipping two coins of different biases, $p \neq q$.

We note that any machine with a single state is necessarily exact since the observer is synchronized before observing any output. However, the synchronization question in this case is moot. Thus, when discussing exact or nonexact machines, we will always assume $N \geq 2$. Also, since any finite word $w \in \mathcal{L}(M)$ is contained in a.e. infinite sequence $\overrightarrow{x}$ an ϵ-machine M generates, we know

that a machine M is exact if (and only if) it has some sync word w of finite length.

One final important quantity to monitor during synchronization is the observer's average uncertainty in the machine state after seeing a length-L block of output.

Definition 9. *The observer's* average state uncertainty *at time L is:*

$$\begin{aligned}\mathcal{U}(L) &\equiv H[\mathcal{S}_L|\overrightarrow{X}^L] \\ &= \sum_{\{\overrightarrow{x}^L\}} \Pr(\overrightarrow{x}^L) u(\overrightarrow{x}^L) \ .\end{aligned}$$

C. Prediction

A process's intrinsic randomness is measured by its entropy rate and that, in turn, determines how well an observer can predict its behavior.

Definition 10. *The* block entropy $H(L)$ *for a stationary process $\mathcal{P}$ is:*

$$\begin{aligned}H(L) &\equiv H[\overrightarrow{X}^L] \\ &= -\sum_{\{\overrightarrow{x}^L\}} \Pr(\overrightarrow{x}^L) \log_2 \Pr(\overrightarrow{x}_L) \ .\end{aligned}$$

Definition 11. *The* entropy rate h_μ *is the asymptotic average entropy per symbol:*

$$\begin{aligned}h_\mu &\equiv \lim_{L\to\infty} \frac{H(L)}{L} \\ &= \lim_{L\to\infty} H[X_L|\overrightarrow{X}^L] \ .\end{aligned}$$

Definition 12. *Its* length-L approximation *is:*

$$\begin{aligned}h_\mu(L) &\equiv H(L) - H(L-1) \\ &= H[X_{L-1}|\overrightarrow{X}^{L-1}] \ .\end{aligned}$$

That is, $h_\mu(L)$ is the observer's average uncertainty in the next symbol to be generated after observing the first $L-1$ symbols.

For any stationary process, $h_\mu(L)$ monotonically decreases to the limit h_μ [4]. However, the form of convergence depends on the process. The lower the value of h_μ a process has, the better an observer's predictions of the process will be asymptotically. The faster $h_\mu(L)$ converges to h_μ, the faster an observer's predictions will reach this optimal asymptotic level. Since we are often interested in making predictions after only a finite sequence of observations, the source's true entropy rate h_μ, as well as the rate of convergence of $h_\mu(L)$ to h_μ, are both important properties.

Now, for an ϵ-machine, an observer's prediction of the next output symbol is a direct function of the probability distribution over machine states induced by the previously observed symbols. That is,

$$\begin{aligned}&\Pr(X_L = x|\overrightarrow{X}^L = \overrightarrow{x}^L) \\ &\qquad = \sum_k \Pr(x|\sigma_k) \Pr(\mathcal{S}_L = \sigma_k|\overrightarrow{X}^L = \overrightarrow{x}^L) \ .\end{aligned}$$

Hence, the better an observer knows the machine state at the current time, the better it can predict the next symbol the machine generates. And, on average, the closer $\mathcal{U}(L)$ is to 0, the closer $h_\mu(L)$ is to h_μ. Therefore, the rate of convergence of $h_\mu(L)$ to h_μ for an ϵ-machine is closely related to the average rate of synchronization. This is one of the primary motivations for studying the synchronization problem.

III. AN INTUITIVE PICTURE

In this section we present an intuitive picture of the synchronization process and use it to derive a formula for the conditional state distribution $\phi(w)$. The basic idea is illustrated schematically in Fig. 4 for a hypothetical 5-state machine.

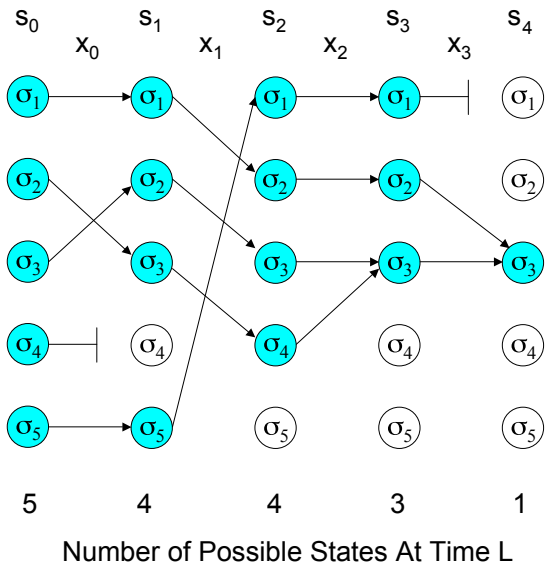

FIG. 4: Synchronization illustrated for a 5-state machine.

Initially, the observer does not know the machine state $\mathcal{S}_0$, so all five states $\{\sigma_1, \sigma_2, \sigma_3, \sigma_4, \sigma_5\}$ are possible. After seeing the first symbol x_0, there are only four possibilities for $\mathcal{S}_1$—$\{\sigma_1, \sigma_2, \sigma_3, \sigma_5\}$—since only four of the five states may generate this symbol. After seeing the second symbol x_1, a different set of four states is possible—$\{\sigma_1, \sigma_2, \sigma_3, \sigma_4\}$. After seeing the third symbol x_2, there are only three possibilities $\{\sigma_1, \sigma_2, \sigma_3\}$ for $\mathcal{S}_3$, since two of the state paths merge on seeing the third symbol. Finally, after seeing the fourth symbol x_3, two more state paths merge and another dies, so there is only one possibility $\{\sigma_3\}$ for $\mathcal{S}_4$. The observer has synchronized.

The transition function $\delta(\sigma_k, x)$ is defined by the relation $\sigma_k \stackrel{x}{\to} \delta(\sigma_k, x)$ and the word transition function $\delta(\sigma_k, w)$ by the relation $\sigma_k \stackrel{w}{\to} \delta(\sigma_k, w)$. In general, for each possible initial state σ_k and each $\overrightarrow{x}^L$ there is a state path p^k following σ_k under $\overrightarrow{x}^L$. That is, $p^k = p^k_0, p^k_1, \ldots, p^k_L$, where $p^k_0 = \sigma_k$, $p^k_1 = \delta(p^k_0, x_0)$, $p^k_2 = \delta(p^k_1, x_1)$, and so on. An observer synchronizes exactly once all, but one, of these paths have either merged or died.

If a machine is not exactly synchronizable, then it is impossible for all paths to merge or die and, at any finite time L, there are at least two possible nonmerged paths remaining. However, it is still possible for an observer to synchronize to such a machine asymptotically. To understand how this happens we need to know the relative probabilities of being in each of the possible remaining states at a given time. In general, the probability of starting in state σ_k and generating the word $\overrightarrow{x}^L$ is:

$$\begin{aligned}\Pr(p^k) &\equiv \Pr(\mathcal{S}_0 = \sigma_k, \overrightarrow{X}^L = \overrightarrow{x}^L) \\ &= \pi_k \cdot \Pr(\overrightarrow{x}^L | \sigma_k) \ .\end{aligned}$$

These probabilities will be exactly 0 if and only if the path p^k dies by the L_{th} symbol. Typically, however, all these probabilities decay—in fact, decay exponentially fast—as $L \to \infty$. For nonexact synchronization, we are concerned not with absolute path probabilities, but with their relative or normalized probabilities. The probability of ending up in state σ_j at time L is simply the sum of the normalized probabilities of all paths ending up in state σ_j. That is, for any word $w = \overrightarrow{x}^L$ we have:

$$\begin{aligned}\phi(w)_j &\equiv \Pr(\mathcal{S}_L = \sigma_j | \overrightarrow{X}^L = w) \\ &= \frac{\Pr(\mathcal{S}_L = \sigma_j, \overrightarrow{X}^L = w)}{\Pr(\overrightarrow{X}^L = w)} \\ &= \frac{\sum_k \pi_k \cdot \Pr(w|\sigma_k) \cdot I(w,k,j)}{\sum_i \pi_i \cdot \Pr(w|\sigma_i)} \\ &= \frac{\sum_k \Pr(p^k) \cdot I(w,k,j)}{\sum_i \Pr(p^i)} \ . \end{aligned} \tag{1}$$

For nonexact asymptotic synchronization, then, the important quantities to consider are the relative probabilities of all paths that never merge or die. If a nonexact machine is asymptotically synchronizable, then for a.e. $\overrightarrow{x}$ there must be some state σ_k such that the ratio of the path probabilities:

$$\frac{\Pr(p^k(\overrightarrow{x}^L))}{\Pr(p^j(\overrightarrow{x}^L))} \to \infty \ ,$$

as $L \to \infty$, for any path p^j which does not eventually merge with p^k or die. Since π_k/π_j is bounded for all states σ_k and σ_j, the initial state is unimportant for asymptotic synchronization. The question is whether, on average, the transition probabilities for one path are greater than those of the other. If, on average, the transition probabilities for path p^k are c (> 1) times as likely as the transition probabilities for path p^j then, for large L, $\Pr(p^k)/\Pr(p^j) \sim c^L$. Intuitively, this is why synchronization occurs exponentially fast. Establishing this, as we will see, requires some care, however.

IV. THE ENTROPY RATE FORMULA

In this section we derive a formula for the entropy rate of a finite-state ϵ-machine. Although an analogous expression has been previously established in similar contexts (see, e.g., Ref. [5]), we provide a derivation as well for completeness. The proof presented here is also somewhat simpler than the original in Ref. [5]. A proof quite similar to ours for unifilar Moore hidden Markov models (as opposed to the edge-label or Mealy models we use) is given in Ref. [6].

Proposition 1. *For any ϵ-machine M,*

$$\begin{aligned}h_\mu &= H[X_0|\mathcal{S}_0] \\ &\equiv \sum_k \pi_k h_k \ , \end{aligned} \tag{2}$$

where $h_k = H[X_0|\mathcal{S}_0 = \sigma_k]$.

Proof. We establish the bounds from above and below separately.

Upper bound: $h_\mu \leq H[X_0|\mathcal{S}_0]$. We calculate directly:

$$\begin{aligned}h_\mu &\equiv \lim_{L\to\infty} \frac{H[\overrightarrow{X}^L]}{L} \\ &\leq \lim_{L\to\infty} \frac{H[\mathcal{S}_0, \overrightarrow{X}^L]}{L} \\ &\stackrel{(a)}{=} \lim_{L\to\infty} \frac{H[\mathcal{S}_0] + \sum_{i=0}^{L-1} H[X_i|\mathcal{S}_0, \overrightarrow{X}^i]}{L} \\ &\stackrel{(b)}{=} \lim_{L\to\infty} \frac{H[\mathcal{S}_0] + \sum_{i=0}^{L-1} H[X_i|\mathcal{S}_i]}{L} \\ &\stackrel{(c)}{=} \lim_{L\to\infty} \frac{H[\mathcal{S}_0] + H[X_0|\mathcal{S}_0] \cdot L}{L} \\ &= H[X_0|\mathcal{S}_0] \ , \end{aligned}$$

where step (a) follows from the chain rule, step (b) from unifilarity, and step (c) from stationarity.

Lower bound: $h_\mu \geq H[X_0|\mathcal{S}_0]$. We have:

$$\begin{aligned}h_\mu &\equiv \lim_{L\to\infty} \frac{H[\overrightarrow{X}^L]}{L} \\ &\geq \lim_{L\to\infty} \frac{H[\overrightarrow{X}^L|\mathcal{S}_0]}{L} \\ &\stackrel{(a)}{=} \lim_{L\to\infty} \frac{\sum_{i=0}^{L-1} H[X_i|\mathcal{S}_0, \overrightarrow{X}^i]}{L} \\ &\stackrel{(b)}{=} \lim_{L\to\infty} \frac{\sum_{i=0}^{L-1} H[X_i|\mathcal{S}_i]}{L} \\ &\stackrel{(c)}{=} \lim_{L\to\infty} \frac{H[X_0|\mathcal{S}_0] \cdot L}{L} \\ &= H[X_0|\mathcal{S}_0] \ , \end{aligned}$$

where again step (a) follows from the chain rule, step (b) from unifilarity, and step (c) from stationarity. □

V. AVERAGED ASYMPTOTIC SYNCHRONIZATION

The entropy rate formula says that (on average) an observer predicts asymptotically just as well as if it knew

the machine state exactly:

$$\begin{aligned} h_\mu &= \lim_{L\to\infty} H[X_L|\overrightarrow{X}^L] \\ &= H[X_0|\mathcal{S}_0] \ . \end{aligned}$$

Intuitively, this suggests that an observer's average uncertainty $\mathcal{U}(L)$ in the machine state must vanish asymptotically. That is, we should have:

$$\lim_{L\to\infty} \mathcal{U}(L) = 0 \ .$$

These ideas are made rigorous below with a convexity argument.

The following notation will be used:

- Let $p_k \equiv \Pr(X_0|\mathcal{S}_0 = \sigma_k)$ and $p_w \equiv \Pr(X_0|\mathcal{S}_0 \sim \phi(w))$.
- Let $h_k \equiv H[p_k]$ (as above), $h_w \equiv H[p_w]$, and $\tilde{h}_w \equiv \sum_k \phi(w)_k h_k$.
- Let $A_{\epsilon,L} \equiv \{w \in \mathcal{L}_L(M) : u(w) < \epsilon\}$ and $A^c_{\epsilon,L} \equiv \mathcal{L}_L(M)/A_{\epsilon,L}$, the complement of $A_{\epsilon,L}$.

We note that for any word w:

$$p_w = \sum_k p_k \phi(w)_k \ , \tag{3}$$

and, hence, by the concavity of the entropy function $H[\cdot]$:

$$\begin{aligned} h_w &= H[p_w] \\ &= H\left[\sum_k p_k \phi(w)_k\right] \\ &\geq \sum_k \phi(w)_k H[p_k] \\ &= \sum_k \phi(w)_k h_k \\ &= \tilde{h}_w \ . \end{aligned} \tag{4}$$

Also, for any length L:

$$\begin{aligned} h_\mu(L+1) &= H[X_L|\overrightarrow{X}^L] \\ &= \sum_{w\in\mathcal{L}_L(M)} \Pr(w) h_w \ , \end{aligned} \tag{5}$$

and

$$\begin{aligned} h_\mu &= \sum_k \pi_k h_k \\ &= \sum_k \left(\sum_{w\in\mathcal{L}_L(M)} \Pr(w)\phi(w)_k \right) h_k \\ &= \sum_{w\in\mathcal{L}_L(M)} \Pr(w) \sum_k \phi(w)_k h_k \\ &= \sum_{w\in\mathcal{L}_L(M)} \Pr(w) \tilde{h}_w \ . \end{aligned} \tag{6}$$

Proposition 2. *For any finite-state ϵ-machine M:*

$$\lim_{L\to\infty} \mathcal{U}(L) = 0 \ . \tag{7}$$

Proof. We first prove the statement under the assumption that M has a minimum distinguishing length $L^* = 1$. We then use this result to establish the general case.

Case (i): M has minimum distinguishing length $L^* = 1$. The proof is by contradiction. If $\mathcal{U}(L) \not\to 0$, then there must be some $\epsilon > 0$ for which $\Pr(A^c_{\epsilon,L}) \not\to 0$. Hence, there exists some $\delta > 0$ and a subsequence $(L_i)_{i=1}^\infty$ of the Ls such that $\Pr(A^c_{\epsilon,L_i}) \geq \delta$, for all i.

Let Δ be the unit simplex in $\mathbb{R}^N$:

$$\Delta = \left\{ \phi \in \mathbb{R}^N : \sum_k \phi_k = 1 \text{ and } \phi_k \geq 0, \text{ for all } k \right\} \ ,$$

and let:

$$\Delta_\epsilon = \{\phi \in \Delta : H[\phi] \geq \epsilon\} \ .$$

Define $f : \Delta_\epsilon \to \mathbb{R}$ by:

$$f(\phi) = H\left[\sum_k \phi_k p_k\right] - \sum_k \phi_k H[p_k] \ ,$$

so that, for any word w, $f(\phi(w)) = h_w - \tilde{h}_w$.

Then, with respect to $\|\cdot\|_1$, $f(\phi)$ is a continuous function on Δ_ϵ and Δ_ϵ is a compact set. Therefore, we know f obtains its minimum f^* at some point $\phi^* \in \Delta_\epsilon$.

Since M has a minimum distinguishing length $L^* = 1$ and the entropy function $H[\cdot]$ is strictly concave, we know $f(\phi) > 0$ for all $\phi \in \Delta_\epsilon$. In particular, $f(\phi^*) = f^* > 0$.

Hence, for each i we have:

$$\begin{aligned} & h_\mu(L_i+1) - h_\mu \\ &\overset{(a)}{=} \sum_{w\in\mathcal{L}_{L_i}(M)} \Pr(w) h_w - \sum_{w\in\mathcal{L}_{L_i}(M)} \Pr(w)\tilde{h}_w \\ &= \sum_{w\in\mathcal{L}_{L_i}(M)} \Pr(w) \cdot (h_w - \tilde{h}_w) \\ &\overset{(b)}{\geq} \sum_{w\in A^c_{\epsilon,L_i}} \Pr(w) \cdot (h_w - \tilde{h}_w) \\ &\geq \Pr(A^c_{\epsilon,L_i}) \cdot f^* \\ &\geq \delta f^* \ . \end{aligned} \tag{8}$$

Step (a) follows from Eqs. (5) and (6) and step (b) follows from Eq. (4). Equation (8) implies that $h_\mu(L) \not\to h_\mu$, which is a contradiction. Hence, we know $\lim_{L\to\infty} \mathcal{U}(L) = 0$.

Case (ii): M has minimum distinguishing length $L^* > 1$. Take $n \geq L^*$ and relatively prime to the period p of M's graph, so that M^n is an ϵ-machine with a minimum distinguishing length of 1. Let Y_L be the RV for the L_{th} output symbol generated by the machine M^n, let R_L be

the RV for M^n's L_{th} state, and let $\mathcal{V}(L) = H[R_L|\overrightarrow{Y}^L]$. Note that for any L:

$$\mathcal{V}(L) = \mathcal{U}(nL). \tag{9}$$

Now, for a contradiction assume $\lim_{L\to\infty}\mathcal{U}(L) \neq 0$. Then, since $\mathcal{U}(L)$ is monotonically decreasing, we know there exists some $\epsilon > 0$ such that $\mathcal{U}(L) \geq \epsilon$, for all L. Thus, by Eq. (9), we know that $\mathcal{V}(L) \geq \epsilon$ for all L as well, so $\mathcal{V}(L) \not\to 0$. However, since M^n has a minimum distinguishing length of 1, by case (i) above we know that $\mathcal{V}(L)$ must go to zero. This contradiction implies that $\lim_{L\to\infty}\mathcal{U}(L) = 0$. □

VI. THE NONEXACT MACHINE SYNCHRONIZATION THEOREM

In this section we prove our primary result, the Nonexact Machine Synchronization Theorem. This extends the weak asymptotic synchronization result of Sec. V to show that synchronization occurs exponentially fast for nonexact machines. The statement is quite analogous to the Exact Machine Synchronization Theorem given in Ref. [1]. Essentially, it says that, except on a set of words $\overrightarrow{x}^L$ of exponentially small probability, an observer's uncertainty after observing $\overrightarrow{x}^L$ is exponentially small.

The following notation will be used. $\Phi_L \equiv \phi(\overrightarrow{X}^L)$ is the random variable for the belief distribution over states induced by the first length-L word the machine generates, and $\overline{\mathcal{S}}_L$ is the most likely state in Φ_L (if a tie the lowest numbered state is taken). $P_L \equiv \Pr(\overline{\mathcal{S}}_L)$ is the probability of the most likely state in the distribution Φ_L, and $Q_L \equiv \Pr(\text{NOT } \overline{\mathcal{S}}_L)$ is the combined probability of all other states in the distribution Φ_L. For example, if $\Phi_L = (0.2, 0.7, 0.1)$, then $\overline{\mathcal{S}}_L = \sigma_2$, $P_L = 0.7$, and $Q_L = 0.3$. Realizations are denoted ϕ_L, $\overline{s}_L$, p_L, and q_L, respectively. We also define $U_L = H[\Phi_L]$ and $u_L = H[\phi_L]$.

Theorem 1. *(Nonexact Machine Synchronization Theorem) For any nonexact ϵ-machine M,*

1. *There exist constants $K_1 > 0$ and $0 < \alpha_1 < 1$ such that:*

$$\Pr(Q_L > \alpha_1^L) \leq K_1\alpha_1^L \ , \ \textit{for all } L \in \mathbb{N}.$$

2. *There exist constants $K_2 > 0$ and $0 < \alpha_2 < 1$ such that:*

$$\Pr(U_L > \alpha_2^L) \leq K_2\alpha_2^L \ , \ \textit{for all } L \in \mathbb{N}.$$

The proof strategy is as follows. We first take a power machine M^n of the machine M with $\mathcal{U}(n) = \epsilon \ll 1$, and prove the theorem for the power machine. We then use the exponential convergence of the power machine to establish the theorem in general with a subsequence-type argument.

The following lemma on large deviations of Markov chains will be critical.

Lemma 1. *Let $Z_0, Z_1, ...$ be a finite-state, irreducible Markov chain, with state set $R = \{r_1, ..., r_n\}$ and equilibrium distribution $\rho = (\rho_1, ..., \rho_n)$. Let $F : R \to \mathbb{R}$, $Y_L = F(Z_L)$, and $\overline{Y}_L = \frac{1}{L}(Y_0 + ... + Y_{L-1})$. Define $\mu_F = \mathbb{E}_\rho(F) = \sum_k \rho_k F(r_k)$. Then, for any $\epsilon > 0$, there exist constants $K > 0$ and $0 < \alpha < 1$ such that, for any state r_k:*

$$\Pr\left(|\overline{Y}_L - \mu_F| > \epsilon | \mathcal{S}_0 = r_k\right) \leq K\alpha^L \ , \ \textit{for all } L \in \mathbb{N} \ .$$

Proof. A similar statement (with more explicit values of the constants) is given in Ref. [7] for a general class of Markov chains, which includes all finite-state, irreducible, aperiodic chains. The result stated here follows directly for finite-state, irreducible, aperiodic chains, and can be extended to the periodic case by considering length p-blocks, where p is the chain's period. □

Remark. *Note that since the deviation bound holds conditionally on any initial state r_k, it also holds conditionally on any distribution over the initial state by linearity. In particular, we apply this lemma assuming $Z_0 \sim \rho$.*

Let us denote:

$$\begin{aligned}
\Pr(x, \sigma_k) &= \Pr(\mathcal{S}_0 = \sigma_k, X_0 = x) \ , \\
\Pr(x|\sigma_k) &= \Pr(X_0 = x|\mathcal{S}_0 = \sigma_k) \ , \\
\Pr(\sigma_k|x) &= \Pr(\mathcal{S}_0 = \sigma_k|X_0 = x) \ , \text{ and} \\
\sigma_{\max,x} &= \text{argmax } \Pr(\sigma_k|x) \ ,
\end{aligned}$$

where again the lowest numbered state is chosen in the case of a tie for $\sigma_{\max,x}$. Also, for any x and σ_j with $\Pr(x|\sigma_j) > 0$, let us define:

$$\begin{aligned}
\boldsymbol{\mathcal{S}}_{x,j} &= \{\sigma_k \in \boldsymbol{\mathcal{S}} : \Pr(x|\sigma_k) > 0 \ , \ \delta(\sigma_k, x) \neq \delta(\sigma_j, x)\} \ , \\
g(x, \sigma_j) &= \max_{\sigma_k \in \boldsymbol{\mathcal{S}}_{x,j}} \Pr(x|\sigma_k) \ , \text{ and} \\
f(x, \sigma_j) &= \max_{\sigma_k \in \boldsymbol{\mathcal{S}}_{x,j}} \Pr(\sigma_k|x) \ .
\end{aligned}$$

Note that $g(x, \sigma_j)$ and $f(x, \sigma_j)$ are both always strictly positive for nonexact ϵ-machines. And, also, that for any joint length-L realization $(\overrightarrow{x}^L, \overrightarrow{s}^L)$:

$$\frac{p_L}{q_L} \geq \frac{\Pr(s_0)}{\Pr(\text{NOT } s_0)} \prod_{i=0}^{L-1} \frac{\Pr(x_i|s_i)}{g(x_i, s_i)} \ , \tag{10}$$

by Eq. (1). Here, $\Pr(s_0) = \pi_k$ is the stationary probability of the state $s_0 = \sigma_k$ and $\Pr(\text{NOT } s_0) = 1 - \Pr(s_0)$.

Using Lemma 1 we now prove our desired theorem under the (relatively strong) assumption that:

$$\mathbb{E}_{\pi_{edge}}\left\{\log_2\left(\frac{\Pr(X_0|\mathcal{S}_0)}{g(X_0, \mathcal{S}_0)}\right)\right\} > 0 \ . \tag{11}$$

This assumption will later be satisfied for some power machine M^n.

Lemma 2. *Let M be a nonexact ϵ-machine satisfying Eq. (11). Then* :

1. *There exist constants $K_1 > 0$ and $0 < \alpha_1 < 1$ such that:*

$$\Pr(Q_L > \alpha_1^L) \leq K_1 \alpha_1^L \ , \ \textit{for all } L \in \mathbb{N}.$$

2. *There exist constants $K_2 > 0$ and $0 < \alpha_2 < 1$ such that:*

$$\Pr(U_L > \alpha_2^L) \leq K_2 \alpha_2^L \ , \ \textit{for all } L \in \mathbb{N}.$$

Proof. We first prove Claim 1 and then use this to show Claim 2.

Proof of Claim 1: Consider the edge-label Markov process $\mathcal{P}_{edge}$ generated by the edge machine M_{edge}. Let $Z_L = (X_L, \mathcal{S}_L)$ denote the RV for the L_{th} M_{edge}-state and let:

$$\begin{aligned} Y_L &= F(Z_L) \\ &= \log_2 \left(\frac{\Pr(X_L|\mathcal{S}_L)}{g(X_L, \mathcal{S}_L)} \right) \ . \end{aligned}$$

We assume, of course, that $(X_0, \mathcal{S}_0) \sim \pi_{edge}$ or, equivalently, $\mathcal{S}_0 \sim \pi$.

By hypothesis, $\mu_F = \mathbb{E}_{\pi_{edge}}(F) = C > 0$. Take $\epsilon = C/2$. By Lemma 1, there exist constants $B_1 > 0$ and $0 < \eta_1 < 1$ such that:

$$\Pr(|\overline{Y}_L - \mu_F| > \epsilon) \leq B_1 \eta_1^L \ , \ \text{for all } L \in \mathbb{N} \ .$$

Thus, for any L:

$$\begin{aligned} \Pr\left(\sum_{i=0}^{L-1} \log_2 \left(\frac{\Pr(X_i|\mathcal{S}_i)}{g(X_i, \mathcal{S}_i)} \right) < \frac{C}{2} L \right) &= \Pr(\overline{Y}_L < C/2) \\ &= \Pr(\overline{Y}_L < \mu_F - \epsilon) \\ &\leq \Pr(|\overline{Y}_L - \mu_F| > \epsilon) \\ &\leq B_1 \eta_1^L \ . \end{aligned}$$

Now, let $\overrightarrow{z}^L = (\overrightarrow{x}^L, \overrightarrow{s}^L)$ be any *typical sequence*, i.e.:

$$\sum_{i=0}^{L-1} \log_2 \left(\frac{\Pr(x_i|x_i)}{g(x_i, x_i)} \right) \geq \frac{C}{2} L.$$

Taking logarithms of Eq. (10) we find:

$$\begin{aligned} \log_2 \left(\frac{p_L}{q_L} \right) &\geq \log_2 \left(\frac{\Pr(s_0)}{\Pr(\text{NOT } s_0)} \right) + \sum_{i=0}^{L-1} \log_2 \left(\frac{\Pr(x_i|s_i)}{g(x_i, s_i)} \right) \\ &\geq \beta + \frac{C}{2} L \ , \end{aligned}$$

where $\beta \equiv \min_k \log_2 \left(\frac{\Pr(\mathcal{S}_0 = \sigma_k)}{\Pr(\mathcal{S}_0 \neq \sigma_k)} \right)$. Or, equivalently:

$$\frac{p_L}{q_L} \geq 2^\beta \cdot 2^{\frac{C}{2} L} = B_2 \eta_2^L \ ,$$

where $B_2 \equiv 2^\beta > 0$ and $\eta_2 \equiv 2^{C/2} > 1$. Thus:

$$q_L \leq \frac{q_L}{p_L} \leq B_3 \eta_3^L,$$

where $B_3 \equiv 1/B_2 > 0$ and $\eta_3 \equiv 1/\eta_2 < 1$. Since this holds for any typical sequence $(\overrightarrow{x}^L, \overrightarrow{s}^L)$ we have, for each L:

$$\Pr(Q_L > B_3 \eta_3^L) \leq B_1 \eta_1^L.$$

And, therefore, for any $1 > \alpha_1 > \max\{\eta_1, \eta_3\}$ there exists some $K_1 = K_1(\alpha_1)$ sufficiently large that:

$$\Pr(Q_L > \alpha_1^L) \leq K_1 \alpha_1^L \ , \ \text{for all } L \in \mathbb{N} \ .$$

Proof of Claim 2: By Claim 1 we know there exist constants $K_1 > 0$ and $0 < \alpha_1 < 1$ such that:

$$\Pr(Q_L > \alpha_1^L) \leq K_1 \alpha_1^L \ , \ \text{for all } L \in \mathbb{N} \ .$$

Let us define:

$$V_L^+ = \{ \overrightarrow{x}^L : q_L > \alpha_1^L \} \text{ and } V_L^- = \{ \overrightarrow{x}^L : q_L \leq \alpha_1^L \} \ .$$

Take L_1 sufficiently large that $1 - \alpha_1^L \geq 1/2$, for all $L \geq L_1$. Note that the first-order Taylor expansion about $x = 1$ of $\log_2(1 - \alpha_1^L)$ is $-\log_2(e)\alpha_1^L + O(\alpha_1^{2L}) \approx -1.44\alpha_1^L + O(\alpha_1^{2L})$. Thus, there exists some $L_2 \in \mathbb{N}$ such that $|\log_2(1 - \alpha_1^L)| \leq 2\alpha_1^L$ for all $L \geq L_2$. Take $L_0 = \max\{L_1, L_2\}$.

Then, for any $L \geq L_0$ and any $\overrightarrow{x}^L \in V_L^-$, we have:

$$\begin{aligned} & H[\mathcal{S}_L|\overrightarrow{x}^L] \\ & \overset{(a)}{\leq} H\left[\left(1 - \alpha_1^L, \frac{\alpha_1^L}{N-1}, \ldots, \frac{\alpha_1^L}{N-1} \right) \right] \\ & = -\left[(1 - \alpha_1^L) \log_2(1 - \alpha_1^L) + \alpha_1^L \log_2 \left(\frac{\alpha_1^L}{N-1} \right) \right] \\ & = -(1 - \alpha_1^L) \log_2(1 - \alpha_1^L) \\ & \qquad - \alpha_1^L L \log_2(\alpha_1) + \alpha_1^L \log_2(N-1) \\ & \overset{(b)}{\leq} (1 - \alpha_1^L) 2\alpha_1^L - \alpha_1^L L \log_2(\alpha_1) + \alpha_1^L \log_2(N-1) \\ & \leq L C_1 \alpha_1^L \\ & \leq C_2 \alpha^L \ , \end{aligned} \tag{12}$$

where $C_1 \equiv 2 - \log_2(\alpha_1) + \log_2(N-1) > 0$, step (a) follows from the fact that $1 - \alpha_1^L \geq 1/2$ for $L \geq L_1$, and step (b) follows from the Taylor expansion bound on $|\log_2(1 - \alpha_1^L)|$ for $L \geq L_2$. In the last line, α may be chosen as any real number in the interval $(\alpha_1, 1)$ and $C_2 = C_2(\alpha)$ is chosen sufficiently large to ensure the last inequality holds for all $L \geq L_0$.

Equation (12) implies that, for all $L \geq L_0$:

$$\Pr(U_L \leq C_2 \alpha^L) \geq \Pr(V_L^-) \geq 1 - K_1 \alpha_1^L \ .$$

So, we know that, for all $L \geq L_0$:

$$\Pr(U_L > C_2 \alpha^L) \leq K_1 \alpha_1^L \leq K_1 \alpha^L \ .$$

Therefore, for any $\alpha_2 \in (\alpha, 1)$ and L sufficiently large:

$$\Pr(U_L > \alpha_2^L) \leq K_1 \alpha_2^L \ .$$

And, hence, there exists some $K_2 \geq K_1$ such that:

$$\Pr(U_L > \alpha_2^L) \leq K_2 \alpha_2^L \ , \text{ for all } L \in \mathbb{N} \ .$$

□

To establish the theorem in general now, we show that for any machine M there exists a power machine M^n satisfying Eq. (11). To do so requires several additional lemmas.

Lemma 3. *Let M be a nonexact ϵ-machine. Then, for all x and σ_j with $Pr(x|\sigma_j) > 0$:*

$$g(x, \sigma_j) \leq f(x, \sigma_j) \frac{\Pr(x)}{\Pr(\sigma_j)} \lambda^2 \ ,$$

where $\lambda \equiv \max_{i,j} \pi_i / \pi_j$ and $\Pr(\sigma_j)$ and $\Pr(x)$ are the respective stationary probabilities of the state σ_j and symbol x: $\Pr(\sigma_j) = \pi_j$ and $\Pr(x) = \Pr(X_0 = x | \mathcal{S}_0 \sim \pi)$.

Proof. Fix x and σ_j. Take $\sigma_{k_1} \in \boldsymbol{\mathcal{S}}_{x,j}$ such that:

$$\Pr(\sigma_{k_1}|x) / \Pr(\sigma_{k_1}) = \max_{\sigma_k \in \boldsymbol{\mathcal{S}}_{x,j}} \Pr(\sigma_k|x) / \Pr(\sigma_k) \ ,$$

and take $\sigma_{k_2} \in \boldsymbol{\mathcal{S}}_{x,j}$ such that:

$$\Pr(\sigma_{k_2}|x) = \max_{\sigma_k \in \boldsymbol{\mathcal{S}}_{x,j}} \Pr(\sigma_k|x) \ .$$

Then:
$$\begin{aligned} \frac{\Pr(\sigma_{k_2}|x)}{\Pr(\sigma_{k_2})} &\geq \frac{\Pr(\sigma_{k_1}|x)}{\Pr(\sigma_{k_2})} \\ &= \frac{\Pr(\sigma_{k_1}|x)}{\Pr(\sigma_{k_1})} \cdot \frac{\Pr(\sigma_{k_1})}{\Pr(\sigma_{k_2})} \\ &\geq \frac{\Pr(\sigma_{k_1}|x)}{\Pr(\sigma_{k_1})} \cdot 1/\lambda \ , \end{aligned}$$

and
$$\begin{aligned} \frac{\Pr(\sigma_{k_2}|x)}{\Pr(\sigma_j)} &= \frac{\Pr(\sigma_{k_2}|x)}{\Pr(\sigma_{k_2})} \cdot \frac{\Pr(\sigma_{k_2})}{\Pr(\sigma_j)} \\ &\geq \frac{\Pr(\sigma_{k_2}|x)}{\Pr(\sigma_{k_2})} \cdot 1/\lambda \ . \end{aligned}$$

Combining these relations we see that:

$$\frac{\Pr(\sigma_{k_1}|x)}{\Pr(\sigma_{k_1})} \leq \lambda^2 \cdot \frac{\Pr(\sigma_{k_2}|x)}{\Pr(\sigma_j)} \ .$$

And, therefore:
$$\begin{aligned} g(x, \sigma_j) &= \max_{\sigma_k \in \boldsymbol{\mathcal{S}}_{x,j}} \{\Pr(x|\sigma_k)\} \\ &= \max_{\sigma_k \in \boldsymbol{\mathcal{S}}_{x,j}} \left\{ \Pr(\sigma_k|x) \cdot \frac{\Pr(x)}{\Pr(\sigma_k)} \right\} \\ &= \max_{\sigma_k \in \boldsymbol{\mathcal{S}}_{x,j}} \{\Pr(\sigma_k|x)/\Pr(\sigma_k)\} \cdot \Pr(x) \\ &= \frac{\Pr(\sigma_{k1}|x)}{\Pr(\sigma_{k1})} \cdot \Pr(x) \\ &\leq \lambda^2 \cdot \frac{\Pr(\sigma_{k2}|x)}{\Pr(\sigma_j)} \cdot \Pr(x) \\ &= f(x, \sigma_j) \frac{\Pr(x)}{\Pr(\sigma_j)} \lambda^2 \ . \end{aligned}$$

□

Define $A_\epsilon = \mathcal{A}_{\epsilon,1} = \{x \in \mathcal{A} : u(x) < \epsilon\}$.

Lemma 4. *Let M be a nonexact ϵ-machine such that:*

1. $\Pr(A_\epsilon) > 1 - \epsilon$, *for some* $\epsilon < 1/2$, *and*
2. $\Pr(\sigma_{\max,x}|x) / \left(f(x, \sigma_{\max,x}) \lambda^2 \right) \geq 2^{2N^2\lambda^2}$, *for all* $x \in A_\epsilon$.

Then, $\mathbb{E}_{\pi_{edge}} \left\{ \log_2 \left(\frac{\Pr(X_0|\mathcal{S}_0)}{g(X_0, \mathcal{S}_0)} \right) \right\} > 0$.

Proof. Applying Lemma 3 we see:

$$\begin{aligned} &\mathbb{E}_{\pi_{edge}} \left\{ \log_2 \left(\frac{\Pr(X_0|\mathcal{S}_0)}{g(X_0, \mathcal{S}_0)} \right) \right\} \\ &= \sum_{x \in \mathcal{A}} \sum_j \Pr(x, \sigma_j) \log_2 \left(\frac{\Pr(x|\sigma_j)}{g(x, \sigma_j)} \right) \\ &\geq \sum_{x \in \mathcal{A}} \sum_j \Pr(x, \sigma_j) \log_2 \left(\frac{\Pr(\sigma_j|x) \Pr(x) / \Pr(\sigma_j)}{f(x, \sigma_j) \lambda^2 \Pr(x) / \Pr(\sigma_j)} \right) \\ &= \sum_{x \in \mathcal{A}} \Pr(x) \sum_j \Pr(\sigma_j|x) \log_2 \left(\frac{\Pr(\sigma_j|x)}{f(x, \sigma_j) \lambda^2} \right) \\ &= \sum_{x \in A_\epsilon} \Pr(x) \sum_j \Pr(\sigma_j|x) \log_2 \left(\frac{\Pr(\sigma_j|x)}{f(x, \sigma_j) \lambda^2} \right) \\ &\quad + \sum_{x \in A_\epsilon^c} \Pr(x) \sum_j \Pr(\sigma_j|x) \log_2 \left(\frac{\Pr(\sigma_j|x)}{f(x, \sigma_j) \lambda^2} \right) \ . \quad (13) \end{aligned}$$

Now, for any $x \in A_\epsilon^c$ we have:

$$\begin{aligned} &\sum_j \Pr(\sigma_j|x) \log_2 \left(\frac{\Pr(\sigma_j|x)}{f(x, \sigma_j) \lambda^2} \right) \\ &\qquad \geq \sum_j \lambda^2 \left[\frac{\Pr(\sigma_j|x)}{\lambda^2} \log_2 \left(\frac{\Pr(\sigma_j|x)}{\lambda^2} \right) \right] \\ &\qquad \geq \sum_j \lambda^2 \cdot -H \left(\frac{\Pr(\sigma_j|x)}{\lambda^2} \right) \\ &\qquad \geq \sum_j \lambda^2 \cdot (-1) \\ &\qquad = -N\lambda^2 \ , \end{aligned}$$

where $H(\cdot)$ is the binary entropy function. So:

$$\begin{aligned} &\sum_{x \in A_\epsilon^c} \Pr(x) \sum_j \Pr(\sigma_j|x) \log_2 \left(\frac{\Pr(\sigma_j|x)}{f(x, \sigma_j) \lambda^2} \right) \\ &\qquad \geq \Pr(A_\epsilon^c) \cdot -N\lambda^2 \\ &\qquad > -\epsilon N \lambda^2 \ . \quad (14) \end{aligned}$$

Also, if we let $\boldsymbol{\mathcal{S}}^- \equiv \boldsymbol{\mathcal{S}}/\{\sigma_{\max,x}\}$, then for any $x \in A_\epsilon$:

$$\begin{aligned}
&\sum_j \Pr(\sigma_j|x) \log_2 \left(\frac{\Pr(\sigma_j|x)}{f(x,\sigma_j)\lambda^2} \right) \\
&\quad = \Pr(\sigma_{\max,x}|x) \log_2 \left(\frac{\Pr(\sigma_{\max,x}|x)}{f(x,\sigma_{\max,x})\lambda^2} \right) \\
&\qquad + \sum_{\sigma_j \in \boldsymbol{\mathcal{S}}^-} \Pr(\sigma_j|x) \log_2 \left(\frac{\Pr(\sigma_j|x)}{f(x,\sigma_j)\lambda^2} \right) \\
&\quad \geq \frac{1}{N} 2N^2\lambda^2 + \sum_{\sigma_j \in \boldsymbol{\mathcal{S}}^-} \Pr(\sigma_j|x) \log_2 \left(\frac{\Pr(\sigma_j|x)}{f(x,\sigma_j)\lambda^2} \right) \\
&\quad \geq \frac{1}{N} 2N^2\lambda^2 + \sum_{\sigma_j \in \boldsymbol{\mathcal{S}}^-} \lambda^2 \left[\frac{\Pr(\sigma_j|x)}{\lambda^2} \log_2 \left(\frac{\Pr(\sigma_j|x)}{\lambda^2} \right) \right] \\
&\quad \geq 2N\lambda^2 + \sum_{\sigma_j \in \boldsymbol{\mathcal{S}}^-} \lambda^2 \cdot -H\left(\frac{\Pr(\sigma_j|x)}{\lambda^2} \right) \\
&\quad \geq 2N\lambda^2 - N\lambda^2 \\
&\quad = N\lambda^2 \ .
\end{aligned}$$

And, hence:

$$\begin{aligned}
\sum_{x \in A_\epsilon} \Pr(x) \sum_j \Pr(\sigma_j|x) \log_2 \left(\frac{\Pr(\sigma_j|x)}{f(x,\sigma_j)\lambda^2} \right) & \\
\geq \Pr(A_\epsilon) \cdot N\lambda^2 & \\
> (1-\epsilon) N\lambda^2 \ . & \qquad (15)
\end{aligned}$$

Combining Eqs. (13), (14), and (15), we see that:

$$\mathbb{E}\left\{ \log_2 \left(\frac{\Pr(X_0|\mathcal{S}_0)}{g(X_0,\mathcal{S}_0)} \right) \right\} > (1-2\epsilon) N\lambda^2 \equiv C' \ ,$$

where $C' > 0$ for $\epsilon < 1/2$. Since M is not exactly synchronizable, we know $g(x,\sigma_j)$ is always strictly positive, so this expectation must be finite. Hence, there exists some real number $C > C' > 0$ such that:

$$\mathbb{E}\left\{ \log_2 \left(\frac{\Pr(X_0|\mathcal{S}_0)}{g(X_0,\mathcal{S}_0)} \right) \right\} = C \ .$$

□

Remark. *In the above proof we implicitly assumed* $\Pr(x,\sigma_j) \neq 0$ *for all* x *and* j*. The sums for the expectation are, of course, computed only over those* x *and* j *for which* $\Pr(x,\sigma_j) \neq 0$*. Terms involving pairs* (x,σ_j) *with* $\Pr(x,\sigma_j) = 0$ *should be omitted.*

Lemma 5. *For any nonexact* ϵ*-machine* M*, there exists some* $n \in \mathbb{N}$ *such that the power machine* M^n *is an* ϵ*-machine with*

$$\mathbb{E}\left\{ \log_2 \left(\frac{\Pr(Y_0|R_0)}{g(Y_0,R_0)} \right) \right\} > 0 \ ,$$

where Y_L *is the RV for the the* L_{th} *output symbol generated by the machine* M^n *and* R_L *is the RV for the* L_{th} M^n*-state.*

We also denote the alphabet of M^n as $\mathcal{B}$ and the set A_ϵ for the machine M^n as B_ϵ; i.e., $B_\epsilon \sim A_{\epsilon,n}$ for M. We define $\overline{\sigma}(\phi)$ to be the most likely state in a distribution ϕ over the machine states, and $\Pr(\overline{\sigma}(\phi))$ to be the probability of this state in the distribution ϕ. For example, if $\phi = (0.3, 0.1, 0.2, 0.4)$, then $\overline{\sigma}(\phi) = \sigma_4$ and $\Pr(\overline{\sigma}(\phi)) = 0.4$.

Proof. Given any nonexact ϵ-machine M,

1. Take $\overline{\epsilon} = 1/\left(N\lambda^2 2^{2N^2\lambda^2} \right)$.
2. Take ϵ small enough that $\Pr(\overline{\sigma}(\phi)) > 1 - \overline{\epsilon}$ for any state distribution ϕ with $H[\phi] < \epsilon$. (Without loss of generality, we may assume $\epsilon < 1/2$.)
3. For ϵ as above, take n relatively prime to the period p of M's graph and large enough such that $\Pr(A_{\epsilon,n}) > 1 - \epsilon$. (Note that this is possible since $\lim_{L\to\infty} \Pr(A_{\epsilon,L}) = 1$, for all $\epsilon > 0$, since $\lim_{L\to\infty} \mathcal{U}(L) = 0$.)

Then, M^n is an ϵ-machine for the process $\mathcal{P}^n$ and $\Pr(B_\epsilon) = \Pr(A_{\epsilon,n}) > 1 - \epsilon$. Moreover, for all $y \in B_\epsilon$ we have:

$$\begin{aligned}
H[\phi(y)] < \epsilon &\stackrel{(a)}{\Longrightarrow} \Pr(\overline{\sigma}(\phi(y))) > 1 - \overline{\epsilon} \\
&\Longrightarrow f(y,\sigma_{\max,y}) < \overline{\epsilon} \\
&\Longrightarrow \frac{\Pr(\sigma_{\max,y}|y)}{f(y,\sigma_{\max,y})} > \frac{1/N}{\overline{\epsilon}} \\
&\stackrel{(b)}{\Longrightarrow} \frac{\Pr(\sigma_{\max,y}|y)}{f(y,\sigma_{\max,y})} > \lambda^2 2^{2N^2\lambda^2} \\
&\Longrightarrow \frac{\Pr(\sigma_{\max,y}|y)}{f(y,\sigma_{\max,y})\lambda^2} > 2^{2N^2\lambda^2} \ ,
\end{aligned}$$

where step (a) follows from item 2 above and step (b) follows from our choice of $\overline{\epsilon}$. Hence, by Lemma 4:

$$\mathbb{E}\left\{ \log_2 \left(\frac{\Pr(Y_0|R_0)}{g(Y_0,R_0)} \right) \right\} = C > 0 \ .$$

(Note that λ for M is the same as λ for M^n, since M and M^n have the same stationary distribution π.) □

Finally, in order to convert between U_L and Q_L convergence in our theorem we need one last lemma.

Lemma 6. *For any* Φ_L*:*

1. *If* $Q_L \leq 1/2$*, then* $U_L \geq Q_L$*.*
2. *If* $Q_L > 1/2$*, then* $U_L \geq H(1/N)$*, where* $H(\cdot)$ *is the binary entropy function.*

Proof. Note that:

$$\begin{aligned}
U_L &= H[\Phi_L] \\
&\geq H[(1-Q_L, Q_L, 0, \ldots, 0)] \\
&= H(Q_L) \ .
\end{aligned}$$

Since $H(Q_L) \geq Q_L \log_2(1/Q_L)$, we know $H(Q_L) \geq Q_L$ for $Q_L \leq 1/2$. Since $H(Q_L)$ is monotonically decreasing on $[\frac{1}{2}, 1-1/N]$ and Q_L is at most $1-1/N$, we know $H(Q_L) \geq H(1-1/N) = H(1/N)$ for $Q_L > 1/2$. □

Using these lemmas we can now prove the primary result of this section.

Proof. (Nonexact Machine Synchronization Theorem) We first prove Claim 2 of the theorem and then use this to show Claim 1.

Proof of Claim 2: Given any nonexact ϵ-machine M, take a power machine M^n as in Lemma 5 such that:

$$\mathbb{E}\left\{\log_2\left(\frac{\Pr(Y_0|R_0)}{g(Y_0,R_0)}\right)\right\} = C > 0 \ .$$

Denote the random variable U_L for the machine M^n as V_L and the quantity $\mathcal{U}(L)$ for the machine M^n as $\mathcal{V}(L)$. By Lemma 2 we know there exist constants $B_1 > 0$ and $0 < \eta_1 < 1$ such that:

$$\Pr(V_L > \eta_1^L) \leq B_1 \eta_1^L \ , \text{ for all } L \in \mathbb{N}.$$

A proof identical to that of Prop. 3 below then shows there exists some $B_2 > B_1$ such that:

$$\mathcal{V}(L) \leq B_2 \eta_1^L, \text{ for all } L \in N.$$

Or, equivalently:

$$\mathcal{U}(nL) \leq B_2 \eta_1^L \ , \text{ for all } L \in \mathbb{N}.$$

Taking $\eta_2 = \eta_1^{1/n}$ we have:

$$\mathcal{U}(m) \leq B_2 \eta_2^m \ ,$$

for any length m that is an integer multiple of n. Since $\mathcal{U}(m) \leq \log_2(N)$ for any m and is monotonically decreasing, it follows that:

$$\mathcal{U}(m) \leq K \eta_2^m, \text{ for all } m \in \mathbb{N},$$

where $K \equiv \max\{B_2, \log_2(N)\}/\eta_2^n$. Thus, by Markov's inequality, we know that for any $m \in \mathbb{N}$ and $t > 0$:

$$\Pr(U_m > t) \leq \frac{\mathbb{E}U_m}{t} = \frac{\mathcal{U}(m)}{t} \leq \frac{K\eta_2^m}{t} \ .$$

Taking $t = \eta_2^{m/2}$ yields:

$$\Pr(U_m > \alpha^m) \leq K\alpha^m \ ,$$

where $\alpha \equiv \eta_2^{1/2}$.

Proof of Claim 1: By Claim 2 we know there exist constants $K > 0$ and $0 < \alpha < 1$ such that:

$$\Pr(U_L > \alpha^L) \leq K\alpha^L, \text{ for all } L \in \mathbb{N}.$$

Take L_0 large enough that $\alpha^{L_0} < H(1/N)$. Then, for all $L \geq L_0$ we have:

$$\begin{aligned}
\Pr(Q_L > \alpha^L) &= \Pr(\alpha^L < Q_L \leq 1/2) + \Pr(Q_L > 1/2) \\
&\overset{(*)}{\leq} \Pr(U_L > \alpha^L, Q_L \leq 1/2) + \Pr(U_L \geq H(1/N)) \\
&\leq \Pr(U_L > \alpha^L) + \Pr(U_L > \alpha^L) \\
&\leq 2K\alpha^L \ ,
\end{aligned}$$

where step (*) follows from Lemma 6. Hence, for some $\tilde{K} \geq 2K$ we have:

$$\Pr(Q_L > \alpha^L) \leq \tilde{K}\alpha^L \ , \text{ for all } L \in \mathbb{N} \ .$$

□

VII. CONSEQUENCES

As a direct consequence of Thm. 1 we establish exponential convergence results for $\mathcal{U}(L)$ and $h_\mu(L)$ analogous to those in the exact case [1]. We also use Thm. 1 to prove the existence of pointwise almost everywhere (a.e.) exponential synchronization for nonexact machines. This establishes that any ϵ-machine is indeed asymptotically synchronizable in the pointwise sense of Def. 8.

A. Exponential Convergence of $\mathcal{U}(L)$

Proposition 3. *For any nonexact ϵ-machine M there exist constants $K > 0$ and $0 < \alpha < 1$ such that*

$$\mathcal{U}(L) \leq K\alpha^L \ , \text{ for all } L \in \mathbb{N} \ .$$

Proof. Let M be any nonexact ϵ-machine. Then by Thm. 1 there exist constants $C > 0$ and $0 < \alpha < 1$ such that $\Pr(U_L > \alpha^L) \leq C\alpha^L$, for all $L \in \mathbb{N}$. Define:

$$\begin{aligned}
A_L &= \{w \in \mathcal{L}_L(M) : u(w) \leq \alpha^L\} \text{ and} \\
A_L^c &= \mathcal{L}_L(M)/A_L
\end{aligned}$$

Then,

$$\begin{aligned}
\mathcal{U}(L) &= \sum_{w \in \mathcal{L}_L(M)} \Pr(w)u(w) \\
&= \sum_{w \in A_L} \Pr(w)u(w) + \sum_{w \in A_L^c} \Pr(w)u(w) \\
&\leq \Pr(A_L) \cdot \alpha^L + \Pr(A_L^c) \cdot \log_2(N) \\
&\leq 1 \cdot \alpha^L + C\alpha^L \cdot \log_2(N) \\
&= K\alpha^L \ ,
\end{aligned}$$

where $K \equiv 1 + C\log_2(N)$. □

B. Exponential Convergence of $h_\mu(L)$

Proposition 4. *For any nonexact ϵ-machine M, there exist constants $K > 0$ and $0 < \alpha < 1$ such that:*

$$h_\mu(L) - h_\mu \leq K\alpha^L \ , \text{ for all } L \in \mathbb{N} \ .$$

Proof. This follows directly from Prop. 3 and Lemma 7 below. □

Lemma 7. *For any ϵ-machine M and any $L \in \mathbb{N}$:*

$$h_\mu(L+1) - h_\mu \leq \mathcal{U}(L) \ . \tag{16}$$

Proof. Note that:

$$\begin{aligned} H[\overrightarrow{X}^L, X_L, \mathcal{S}_L] &= H[\overrightarrow{X}^L] + H[\mathcal{S}_L|\overrightarrow{X}^L] + H[X_L|\overrightarrow{X}^L, \mathcal{S}_L] \\ &= H[\overrightarrow{X}^L] + H[\mathcal{S}_L|\overrightarrow{X}^L] + H[X_L|\mathcal{S}_L] \\ &= H[\overrightarrow{X}^L] + H[\mathcal{S}_L|\overrightarrow{X}^L] + h_\mu \ , \end{aligned} \tag{17}$$

and also that:

$$\begin{aligned} H[\overrightarrow{X}^L, X_L, \mathcal{S}_L] &= H[\overrightarrow{X}^L] + H[X_L|\overrightarrow{X}^L] \\ &\quad + H[\mathcal{S}_L|\overrightarrow{X}^L, X_L] \ . \end{aligned} \tag{18}$$

Equating the RHS of Eqs. (17) and (18) gives:

$$\begin{aligned} H[\mathcal{S}_L|\overrightarrow{X}^L] + h_\mu &= H[X_L|\overrightarrow{X}^L] + H[\mathcal{S}_L|\overrightarrow{X}^L, X_L] \\ &\geq H[X_L|\overrightarrow{X}^L] \ . \end{aligned} \tag{19}$$

Or, in other words:

$$\mathcal{U}(L) + h_\mu \geq h_\mu(L+1) \ . \tag{20}$$

□

Remark. *If we define the* synchronization *and* predication decay constants*, respectively, as:*

$$\begin{aligned} \alpha_s &= \limsup_{L\to\infty} \mathcal{U}(L)^{1/L} \\ \alpha_p &= \limsup_{L\to\infty} \left(h_\mu(L) - h_\mu\right)^{1/L} \ , \end{aligned}$$

then Lemma 7 also implies that $\alpha_p \leq \alpha_s$. This is to say, the observer's predictions approach their optimal level at least as fast as the observer synchronizes. Since Lemma 7 applies to any ϵ-machine, this statement also holds for any ϵ-machine (exact or nonexact).

C. Pointwise a.e. Asymptotic Synchronization

Proposition 5. *For any nonexact ϵ-machine M there exists some $0 < \alpha < 1$ such that for a.e. $\overrightarrow{x} \in \mathcal{L}_\infty(M)$, there exists $L_0 \in \mathbb{N}$ such that for all $L \geq L_0$,*

$$u(\overrightarrow{x}^L) \leq \alpha^L \ .$$

Proof. Apply the Borel-Cantelli Lemma to Thm. 1. □

VIII. CONCLUSION

We analyzed the process of asymptotic synchronization to nonexact ϵ-machines. Although the treatment is more involved mathematically, the primary results are essentially the same as those for the exact case given in Ref. [1]. An observer's average state uncertainty $\mathcal{U}(L)$ vanishes exponentially fast and, consequently, an observer's average uncertainty in predictions $h_\mu(L)$ converges to the machine's entropy rate h_μ exponentially fast, as well.

We hope to extend the asymptotic synchronization results to more general model classes such as countable-state ϵ-machines or nonunifilar HMMs. We also intend to improve the bounds on the constant α given in the convergence theorems.

Acknowledgments

NT was partially supported by an NSF VIGRE fellowship. This work was partially supported by the Defense Advanced Research Projects Agency (DARPA) Physical Intelligence project via subcontract No. 9060-000709. The views, opinions, and findings contained here are those of the authors and should not be interpreted as representing the official views or policies, either expressed or implied, of the DARPA or the Department of Defense.

[1] N. Travers and J. P. Crutchfield. Exact synchronization for finite-state sources. 2010. Santa Fe Institute Working Paper 10-11-031; arxiv.org:1008.4182 [nlin.CD].

[2] J. P. Crutchfield and D. P. Feldman. Regularities unseen, randomness observed: Levels of entropy convergence. *CHAOS*, 13(1):25–54, 2003. C. J. Ellison, J. R. Mahoney, and J. P. Crutchfield. Prediction, Retrodiction, and the Amount of Information Stored in the Present. J. Stat. Phys. 136(6):1005–1034, 2009; J. R. Mahoney, C. J. Ellison, and J. P. Crutchfield. Information Accessibility and Cryptic Processes. J. Phys. A: Math. Theo. 42:362002, 2009.

[3] J. P. Crutchfield and K. Young. Inferring statistical complexity. *Phys. Rev. Let.*, 63:105–108, 1989.

[4] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley-Interscience, New York, second edition, 2006.

[5] C. E. Shannon and W. Weaver. *The Mathematical Theory of Communication*. University of Illinois Press, Champaign-Urbana, 1962.

[6] J. L. Massey. Markov information sources. In J. K. Skwirzynski, editor, *New Directions in Signal Processing in Communication and Control*, volume E25 of *NATO Advanced Study Institutes Series*, pages 15–26. Noordhoff, Leyden, The Netherlands, 1975.

[7] P. W. Glynn and D. Ormoneit. Hoeffding's inequality for uniformly ergodic Markov chains. *Stat. Prob. Lett.*, 56(2):143–146, 2002.